Name: Achyut Shukla Batch: ML - 1 PRN: 20070122005

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)
(Established under section 3 of the UGC Act 1956)
Re - accredited by NAAC with 'A' Grade
Founder: Prof.Dr. S. B. Mujumdar, M.Sc., Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

Lab Assignment —8

Aim :

Implement KMeans Clustering on a Dataset and conduct analysis of the results.

PART — A

Kmeans Clustering

K-Means clustering is a popular and widely used unsupervised machine learning algorithm for partitioning data points into distinct clusters based on their similarity. It is particularly effective for data clustering when the number of clusters is known in advance or when you want to explore the data's inherent structure. Here's how K-Means clustering works:

1. **Initialization**: The first step in K-Means clustering is to select the number of clusters (K) that you want the algorithm to identify. Then, K initial cluster centroids are chosen. These centroids can be selected randomly or using a more informed strategy.

2. **Assignment**: Each data point is assigned to the cluster whose centroid is closest to it. This assignment is typically based on a distance metric, with Euclidean distance being a common choice.

3. **Update Centroids**: After all data points are assigned to clusters, the centroids of each cluster are recalculated as the mean (average) of the data points in that cluster.

4. **Repeat**: Steps 2 and 3 are iteratively repeated until a stopping criterion is met. Common stopping criteria include a maximum number of iterations or when the centroids no longer change significantly.

5. **Convergence**: When the algorithm converges, the final cluster assignments are determined, and each data point belongs to one of the K clusters.

K-Means clustering has several advantages:

1. **Simplicity**: K-Means is relatively easy to understand and implement.
2. **Efficiency**: It can handle large datasets efficiently, especially with appropriate optimizations.
3. **Scalability**: K-Means works well with both low-dimensional and high-dimensional data.
4. **Deterministic Results**: Given the same initial conditions, K-Means will produce the same results.

However, K-Means clustering has some limitations:

1. Sensitivity to Initial Centroids: The results can be sensitive to the initial placement of cluster centroids, which may lead to suboptimal solutions.
2. Dependence on K: Choosing the right number of clusters (K) can be challenging and may require domain knowledge or trial-and-error.
3. Not Suitable for Non-Globular Clusters**: K-Means tends to perform poorly when clusters are non-spherical or have complex shapes.
4. Outliers: It is sensitive to outliers because it relies on the mean, which is easily influenced by extreme values.

To mitigate some of these limitations, variations of K-Means have been developed, such as K-Means++, which provides a more robust initialization, or hierarchical K-Means, which can be used to discover clusters in a hierarchical structure.

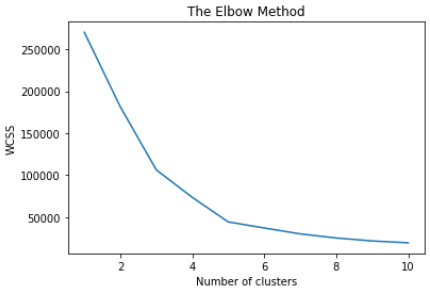
In practice, K-Means clustering is widely used in various applications, including customer segmentation, image compression, document categorization, and more, when the underlying structure of the data is well-suited to K-Means' assumptions.

PART — B

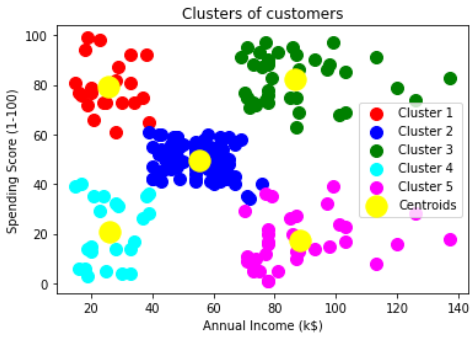
Experiment:

Implement Kmeans Clustering.

1. Elbow Method to Find No of Clusters



2. Cluster Visualization



Inference

Discussion

We have successfully implemented KMeans Clustering visualizing the elbow graph and final clusters.